Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

# OUTLINE OF COURSE AND PRELIMINARY INSTRUCTIONS.

## THE 1928 STATISTICAL CLASS

The question is frequently asked "How can the statistical technique that is found in such books as Mills, Yule, Bowley or taught in the college class room be of use to us in our work?" A good many of our number have either taken a course or two in some college or have made an effort to study some text book in statistics. The development of statistics as a science is of comparatively recent origin. Even 5 years ago it was difficult to get more than an elementary course in the subject in most of our leading universities.

Along with the development of big business and the large corporations has come the science of accounting. In more recent times statistics have been and are being used more and more as an aid to the executive staff in the control of the internal organization as well as in the formation of the future program and policy of the company.

The science of applied statistics has been developed by the statistical staff of these large corporations such as the American Telephone and Telegraph Company, some of the large packers and several large banks. Statistics is like mathematics in that there is considerable difference in the application of the subject matter. We who are workers in one of the largest statistical organizations in the world must also develop an application of statistical theory to our work if we are to keep abreast of the times. It is for this very purpose that we are having our second statistical class and are bringing in 15 men, all statisticians in charge of one or more States. Several of the Washington staff are also planning to be members of this class which will start February 13, 1928 and continue for four weeks. It will be a busy four weeks if we are to cover the ground that should be covered if our course is to be worth while. The course will cost the Bureau more than $100 per day exclusive of salaries of those participating, while the cost including salaries will come to about $10,000. Naturally dividends are expected or this investment would not be made.

With the decentralization of our work it is necessary to put greater dependence upon each man in the states. As individuals we cannot afford to be without the tools with which statisticians in other lines are equipped. Our men are being looked to more and more by State and college authorities for statistical guidance and suggestions. In one State where we have cooperation, no bulletin containing data can receive official approval until it has been passed upon by our State Statistician.

The members of the class will represent a great variety of training and experience. The more work you have had in statistics, mathematics and economics the more you will be expected to contribute to the progress of the class. The course itself will be designed to meet the needs of those who have had no college math or course work in statistics. We are going to start at the very beginning and travel just as far as we can along a well defined pathway. We are to help each other in every way possible as the progress of the class will be only as rapid as the slowest member. For those who feel the need of a little review of High School algebra I would suggest that they obtain the following book and review it carefully. It is recognized as a book that deals with fundamentals and leaves out the non-essentials: Modern Algebra-ninth school year, by Schorling and Clark, World Book Co., 2126 Prairie Avenue, Chicago, Ill.

In order to reduce the clerical work of each individual to a minimum while they are in Washington I am asking that each person do a certain amount of preliminary work before they come in to Washington. This will help us to make the course much more effective and at the same time enable each person to do this work in his own office where he has the help and is used to working. It will require a week or ten days of someone's time and if distributed among several members of the office force will be excellent training for everyone. Detailed instructions concerning this work have been or will be given most of the men personally but will be attached to this letter as a matter of record and for the use of those I may not see.

## STATISTICS APPLIED TO CROP ESTIMATING WORK

Our problems fall largely into two general classes, statistically speaking (1) Sampling (2) Comparisons or Relationships.

Sampling: In only a few cases are we able to go out and by actual count determine how much there is of anything. The Census Bureau does it for agriculture every 5 or 10 years, most of us were familiar with the last Census and some of the problems involved. Some of us are able to obtain a check on our estimates by means of railway shipments, elevator receipts or ginning records. But our first figures are usually estimates based on sample data. Practically every inquiry that we make is a problem in sampling. From sample data we obtain the average yield of wheat or cotton per acre for that sample and then we assume that our sample is a reasonably true one, - containing no bias or error and is fully representative of the "universe" that we are sampling,- and that the average yield for the State as a whole is the same as for our sample.

We do this work so much as a matter of routine that we seldom think of the factors involved in the basic assumptions. Just how dependable is that average after we get it? How much do we dare "lean on it." Would we expect to get the same results if we had a larger sample taken under the same condition? How many reports do we really need with a given inquiry? Do we need as many reports with wheat yield data as with corn yield data, with wheat prices as with apple prices or land value figures? What determines the number of reports needed to give us an average that will not be greatly influenced by the "fluctuation of sampling?"

These are a few of the questions, especially as applied to crop yield data within your State, on which we will turn the search light of statistical theory. Last year we spent two weeks on this work but if the preliminary work is well done prior to starting the class I feel sure that we can complete these phases of the course within a week or ten days and that will leave the remainder of the time to the study of relationships.

Relationships: The most commonly used "relationship" in our work is between the condition of a crop on the first of the month during the growing season and the final yield per acre of that crop at harvest. From 1866 to 1911 the Department made estimates of crop condition during the growing season. In response to popular demand the Department undertook in 1912 to forecast the production of the crop prior to harvest. A method of "condition times par" has been used to indicate the probable yield, which times acreage gives production.

Our study of relationships will start with a study of the relationship of the condition of a crop on a given date with the final yield. First of all is there any relation between the condition of winter wheat on May 1 and the final yield? If the condition is below average what are the probabilities of the yield being less than average? Is the "par" the best measure of this relationship? If there is no relation between condition and final yield in your State for a given crop how shall we go about it to forecast yield per acre? What are some of the other factors that might influence yield that would be available at the time of the forecast? Is it always the weather after the time of the forecast that upsets our early forecasts? These are some of the questions that we will try to answer, each man for his own State.

In case we find some really significant relationships how can they be used in making a forecast? Suppose there are several factors that all have some influence or determination, how can they all be used in making a prediction. This sort of investigation leads us eventually to "correlation" as a means of measuring relationships and when more than one causal factor is used to "Multiple and curvilinear correlation."

There will be ample laboratory and class work which will enable everyone to learn just how the work is done and as much as possible will be done to clear up the underlying philosophy concerning correlation, trends, the use of first differences and logarathisms, coefficients of determination etc.

While the technique used will be familiar to the instructors in charge of the course, the field of sampling as it relates to yield per acre data will be all new territory for everyone. Very little has as yet been done in our division toward the study of factors which determine crop yields. To anyone interested in forecasting crop yields this is a most intriguing field for study and research, an almost unlimited field as each State presents a somewhat different set of factors and relationships than any other.

WORK TO BE COMPLETED PRIOR TO FEBRUARY 10th.

| | | | |
|---|---|---|---|
| References -- | Mills | Pages 61-168 | |
| | | " 548-561 | |
| | | " (516-545) | |
| | U. S. D. A. Bulletin #1480 | Pages 20-28 | |
| | | " 35-37 | Apples |
| | | " 43 | Summary |
| Optional - G. Undy Yule | | Pages 75-105 | Frequency Distribution |
| | | " 106-129 | Averages |
| | | " 133-153 | Measure of Dispersion |

Problem: To form frequency distributions of yield per acre data by means of "tally sheets."

1. Select 6 to 8 of the most important crops in your State for 1927 and one or two other years which represent conditions as much different as 1927 and 1925 as possible. Use regular Field Aid inquiries that have been made in your State. (30 to 40 inquiries should be analyzed) (8 crops for 4 years = 32 samples ) (5 crops with 2 inquiries a year for 4 years = 40 samples)

2. Make out a Tally Sheet by crop reporting districts and for the entire State use a bar graph of the total frequencies. Our stock paper known as (C.E. 1-7) 15" by 25" size with lines about ¼" apart and colums about ½" will prove most satisfactory. Two sheets will be necessary in order to allow from 6 to 8 vertical lines for each crop reporting district and some 20 or 30 lines for the State Summary and bar chart for State frequency table. Six districts can be placed on the front sheet and three districts and State totals on the second sheet. The attached photostat will show how the tally sheet should be made up.

3. Check up and see if your tally sheet includes the following items:
    (1) Show State, crop, year and nature of inquiry at the top of the page, left hand corner. ("New York--Potatoes--1927--Field Aids")
    (2) Rule sheets by districts.
    (3) Show district weights as used with that inquiry.
    (4) Show number of reports by districts.
    (5) Show district averages to one-tenth beyond the decimal.
    (6) Show range, low and high, reported by districts.
    (7) Show class limits on stub to one decimal beyond the unit. usually reported.

4. Number of class intervals should not be less than 8 or 10. (for accuracy) or more than 15 (to avoid too much arithmetic).

(1) Great care should be exercised in seeing that the mid-point of the class interval is a whole number and that a unit that is reported frequently such as 10, 15, 20, or 50, 75, 100, etc., should be made to form the mid-point of a class interval when possible.

(2) The class limits should be definitely defined and equal distance from the mid-point of the class interval. For example

1 bu. class interval, with 5 as a mid-point would have from 4.5 to 5.4 as the class limits.

2 bu. class interval, 5 as mid-point = 4.0 - 5.9 bu.

3 bu. class interval, 5 as mid-point = 3.5 - 6.4 bu.

5 bu. class interval, 10 as mid-point = 7.5 - 12.4 bu.

5. After class interval and mid-points have been determined and entered on the stub 2 or 3 verticle columns will be sufficient for the tally of the reports from each district. Four vertical columns must be left clear between the tally and the beginning of the next district.

6. The tallys for each class interval should be counted 5 at a time, and the number placed in column labeled for each district. ~~llll~~ ~~llll~~ 11 = 12.

7. The State totals for each class interval are obtained by cross addition of district figures and their sum should equal the total number of reports for the State.

8. The use of columns, d, fd and $fd^2$ is explained by both Mills and Yule and briefly as follows: Some central mid-point is selected as the assumed mean or average and a zero placed in the "d" column, and the class intervals numbered consecutively, those above this central mid-point as -1,-2,-3,-4 etc and those below as +1,+2,+3,+4 in the "d" column, these figures being placed in column "d".

Column Fd is found by multiplying the items in "f" by those in "d" giving the correct sign. (Like signs make plus and unlike minus).

Column $fd^2$ is found by multiplying items in column fd by those in d ($d \times fd = fd^2$) (All signs willbe plus).

Chas. F. Sarle.
Agricultural Statistician.